# KnowledgeTB: Leveraging Large Language Models for Enhanced Terminology Extraction over Knowledge Graphs

Konstantinos Chatzitheodorou[1,*,†]

[1] *Aristotle University of Thessaloniki*

## Abstract

The advent of Machine Learning and Large Language Models has revolutionized terminology management, introducing innovative approaches to designing and enriching terminological resources. This paper explores the synergy between Large Language Models and Knowledge Graphs, emphasizing their combined potential to enhance structural design, metadata representation, and data interoperability within specialized domains. Through case studies, we demonstrate how Large Language Models generate domain-specific terminology, ensure linguistic and conceptual precision, and propose metadata by analyzing contextual patterns. Concurrently, we highlight how Knowledge Graphs facilitate the integration of terminological resources into broader ontological frameworks, enabling dynamic updates, enhanced usability, and enriched contextual insights. Our findings propose new strategies for validating terminological resources in terms of ergonomics and usability, offering actionable guidance for experts in terminology, computational linguistics, and knowledge representation. This research contributes to advancing best practices for developing digital terminology systems in the era of Artificial Intelligence.

## Keywords

Terminology Extraction, Terminology Management, Knowledge Graphs, Digital Terminology Systems, Metadata Representation

## 1. Introduction

The advent of advanced computational tools and Machine Learning (ML) techniques has transformed terminology management, shifting it from labor-intensive manual processes to efficient, automated systems [1]. Central to this transformation are Knowledge Graphs (KGs) and Large Language Models (LLMs), which have revolutionized the extraction, organization, and enrichment of terminological resources. This paradigm shift has enabled terminological databases to evolve into dynamic, semantically enriched, and contextually aware assets [2].

LLMs, such as GPT-4 [3] and LLaMA [4], have excelled in natural language understanding and reasoning, facilitating tasks like terminology extraction, validation, and contextual alignment [5]. However, despite their remarkable capabilities, LLMs face significant challenges in handling knowledge-intensive tasks. They often struggle with incomplete or ambiguous information, complex reasoning sequences, and knowledge conflicts stemming from contradictory or outdated data [7, 6]. These conflicts, influenced by their context and type, can profoundly affect model performance, necessitating sophisticated resolution strategies. Additionally, LLMs grapple with issues such as out-of-vocabulary terms, domain shifts, and the computational demands of fine-tuning [5]. As retrieval-augmented LLMs gain wider adoption, addressing these limitations becomes imperative to manage the growing complexity of real-world knowledge tasks.

One promising approach to overcoming these limitations is the integration of LLMs with structured external knowledge sources such as KGs. Platforms like Wikidata [9] and DBpedia [18] exemplify how KGs can complement LLMs, offering structured, interconnected, and multilingual knowledge that strengthens contextual reasoning and enhances semantic understanding [8, 10].

*4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece."*

[*] Corresponding author.

chatzik@itl.auth.gr (K. Chatzitheodorou)
0000-0002-3979-5715 (K. Chatzitheodorou)

In this context, KnowledgeTB introduces a novel framework that synergizes the semantic depth of KGs with the reasoning power of LLMs. Through the integration of these complementary technologies, KnowledgeTB delivers precise, scalable terminology extraction while enriching terms with multilingual metadata and semantic relationships. This hybrid approach adheres to the FAIR principles [11] —Findability, Accessibility, Interoperability, and Reusability— making it suitable for a wide range of applications in specialized domains such as agriculture, healthcare, and tourism.

This paper delves into the methodologies underpinning KnowledgeTB, focusing on its ability to bridge computational linguistics and traditional terminological practices. It explores how the integration of KGs and LLMs optimizes structural design, enhances user-centric analysis, and ensures accurate representation of complex domain-specific knowledge. By doing so, we demonstrate how KnowledgeTB redefines terminology management, facilitating the creation of enriched, interconnected, and globally accessible terminological resources.

## 2. Related work

The integration of LLMs with KGs has been explored across various domains to enhance terminology extraction and knowledge representation. Recent studies have demonstrated the potential of LLMs in specialized machine translation, ontology learning, and keyword extraction. For instance, Kim et al. [13] introduced a methodology that integrates specialized terminology into machine translation models using a term extraction approach based on the Trie Tree algorithm [19]. This method improves the translation accuracy in fields where term consistency is crucial, such as patent translation.

Similarly, LLMs have been applied to ontology learning, where models like GPT-3.5 [21], Llama2-7B [4], and Falcon-7B [20] have been evaluated for tasks such as term typing, taxonomy discovery, and extraction of non-taxonomic relations [14]. These models have proven effective in automating knowledge structuring and improving the usability of ontologies by capturing complex language patterns from large text corpora.

In the context of terminology extraction, LLMs have shown considerable promise in identifying and extracting domain-specific terms with minimal labeled data, aided by innovative prompting strategies and model optimization techniques. The application of LLMs to term extraction tasks, as explored by Tran et al. [15], demonstrates that prompt designs such as text-extractive responses or text-generative responses outperform traditional methods when data is scarce. These advances in LLM-based extraction techniques, when coupled with the structural capabilities of KGs, can significantly improve the creation and management of terminological resources.

Furthermore, LLMs have also proven useful in the context of automatic terminology extraction for various specialized domains. For example, Babaei Giglou et al. [16] have demonstrated how LLMs can be used for ontology learning, which includes term typing, taxonomy discovery, and extracting non-taxonomic relations.

Moreover, the use of external lexical resources, such as WordNet and Wikidata, has proven valuable in enriching under-resourced languages, as shown in McCrae's work [17]. These efforts underscore the growing importance of combining LLMs with KGs to develop digital terminology systems that are both contextually accurate and interoperable across specialized domains.

Our work introduces a hybrid approach that combines the context-aware term extraction capabilities of LLMs with the robust structural framework of KGs. This approach enhances data interoperability, facilitates user-centric metadata representation, and supports multilingual and multidomain applications. By emphasizing the synergy between LLMs and KGs, our research seeks to drive the development of digital terminology systems in an AI-driven era, contributing to best practices in knowledge representation, resource management, and cross-domain applications in diverse linguistic contexts. Unlike previous approaches, KnowledgeTB integrates semantic enrichment and bias mitigation strategies, ensuring domain accuracy and multilingual support.

# 3. Methodology

The development of KnowledgeTB encompasses three core components: terminology extraction using LLMs, semantic enrichment through KGs, and the implementation of multilingual support aligned with the FAIR principles. This methodology ensures the creation of robust, enriched, and universally accessible terminological resources. KnowledgeTB's flexible architecture supports easy integration into terminology management workflows. It can be incorporated into complex workflows within any content management system to automate terminology extraction and linkage. Additionally, it supports standardized data formats, such as CSV, ensuring compatibility with existing terminological systems and databases.

## 3.1. Terminology Extraction Using LLMs

The innovative use of LLMs in KnowledgeTB's terminology extraction is pivotal to its efficacy. Unlike conventional approaches that solely depend on existing lexical databases, KnowledgeTB employs LLMs to dynamically generate domain-specific terms. This adaptive methodology ensures that even emerging and specialized terminologies are efficiently captured.

The prompting process itself follows a methodologically rigorous approach. A typical prompt might resemble the following:

> Extract key terminology related to {domain} from the given text. For each term, provide a brief definition and identify related concepts. Ensure that the terms are domain-specific and exclude colloquial or general language.
>
> Return the response in the following JSON format:

```
{
  "terms": [
    {
      "term": "<extracted_term>",
      "definition": "<brief_definition>",
     "related_concepts": ["<related_concept_1>", "<related_concept_2>"]
    },
    {
      "term": "<extracted_term_2>",
      "definition": "<brief_definition_2>",
     "related_concepts": ["<related_concept_3>", "<related_concept_4>"]
    }
  ]
}
```

Currently, the framework is integrated with OpenAI's GPT-3.5 [21] LLM, leveraging their state-of-the-art language generation capabilities. However, the architecture is designed to be model-agnostic, allowing integration with any other LLM provider. This flexibility ensures that the system can easily adapt to advancements in the field and utilize models best suited for specific tasks or requirements.

For example, given the domain of *Environmental Science*, the prompt may return the following term:

```
{
  "terms": [
    {
      "term": "Afforestation",
      "definition": "The process of planting trees on land that has not
     previously been forested, with the goal of increasing forest cover
```

```
          and enhancing carbon sequestration.",
          "related_concepts": ["Reforestation", "Carbon Sequestration",
          "Ecosystem Restoration"]
        }
      ]
    }
```

### 3.2. Semantic Enrichment Through KGs

The extracted terms undergo semantic enrichment by mapping them to Wikidata entities, which encapsulate multilingual labels, definitions, and interrelated concepts. The term *afforestation* (Q2384419), for example, is enriched with equivalents from over 30 languages and associated with relevant concepts like *forestry* and *planting*. This semantic linking not only facilitates cross-linguistic interoperability but also ensures that the enriched terms are interconnected components of a dynamic and evolving knowledge ecosystem.

The extracted terms undergo semantic enrichment by mapping them to Wikidata entities, which encapsulate multilingual labels, definitions, and interrelated concepts. This semantic enrichment is further complemented by KnowledgeTB's commitment to multilingual accessibility. The term *afforestation*, for example, is linked to its Wikidata entity (Q2384419), which includes multilingual labels such as *afforestation* in French, *forestación* in Spanish, *florestamento* in Portuguese, among many others across 30+ languages. The entity also provides a detailed description of the concept—defined as the establishment of a forest or stand of trees in an area previously lacking tree cover—and includes related subclasses such as *forestry* and *planting*. Additionally, the entity links to external resources, including the *Bibliothèque nationale de France ID* and *GND ID*.

Similarly, the term *ecotourism* is linked to its Wikidata entity (Q187449), which includes multilingual labels such as *turisme ecològic* in Catalan and *Ökotourismus* in German, among others. The entity provides a comprehensive description, defining *ecotourism* as a form of tourism that involves travel to natural attractions and destinations of ecological value, including their living organisms. Additionally, it connects *ecotourism* with facets of protected areas, ecological sites, and national parks, highlighting its role in environmental conservation and sustainable tourism. The entity also links to related resources, such as images that visually represent ecotourism, and connects to other key concepts like *protected area* and *national park*. These connections ensure that the enriched terms are not isolated entries but integral components of a dynamically evolving knowledge ecosystem

Wikidata's multilingual capabilities enable KnowledgeTB to align extracted terms with equivalents in various languages, ensuring that terminological resources remain accessible across linguistic and cultural boundaries. For example, the term *progressive illness* (Q1951525) is linked to its translations in multiple languages, including German (*Progredienz*), Norwegian Bokmål (*progresema malsano*), Turkish (*İlerleyen hastalık*), and Italian (*malattia evolutiva*), promoting inclusivity and fostering global collaboration.

### 3.3. Non-Domain-Specific Terms - Bias

Finally, the response from the model is parsed and filtered to eliminate non-domain-specific terms. The validation phase follows, where each extracted term is cross-referenced with authoritative sources like Wikidata. For instance, if the model identifies the term *carbon sequestration*, it is linked to its Wikidata entity (Q3499912), verifying multilingual labels and related concepts to ensure semantic accuracy.

To address potential issues of bias inherent in LLMs, KnowledgeTB implements a comprehensive bias mitigation strategy. This involves assessing the outputs to identify any language or domain biases generated by the LLMs. Specifically, KnowledgeTB leverages LLMs to cross-check domain consistency between the extracted terms and the domain identified by Wikidata. This process ensures that the system's output remains aligned with the intended domain, preventing the inclusion of terms that may be incorrectly classified or ambiguous. By combining authoritative references with LLM-based validation, the framework upholds high accuracy and relevance in terminological extraction.

### 3.4. Compliance with FAIR Principles

KnowledgeTB strictly adheres to the FAIR principles through a multi-faceted approach. In terms of *Findability*, the framework ensures that each extracted term is associated with persistent and globally unique identifiers from Wikidata, such as QIDs. These identifiers enable unambiguous referencing and facilitate integration with existing digital resources. *Accessibility* is maintained through the use of openly accessible knowledge bases like Wikidata, eliminating licensing restrictions and ensuring that the terminological data remains freely available to users and developers. *Interoperability* is achieved by linking the extracted terms with semantic web technologies, such as RDF and SPARQL endpoints. This alignment ensures compatibility with diverse applications, including environmental modeling and policy frameworks. Finally, *Reusability* is addressed through meticulous documentation and the inclusion of provenance metadata. Each entry is accompanied by detailed information regarding its data sources and extraction methods, facilitating reproducibility and secondary use in academic, governmental, and industrial contexts.

## 4. Results

While the system has not yet undergone exhaustive testing, early evaluations of KnowledgeTB demonstrate its strong performance in terminology extraction and multilingual enrichment. In our preliminary tests, KnowledgeTB successfully identified and enriched a wide array of terms from complex datasets, achieving notable accuracy and semantic depth.

The input text

> "*In the context of sustainable agriculture, terms like carbon footprint and crop rotation are essential for promoting eco-friendly farming practices. In healthcare, the rise of telemedicine has revolutionized patient care, offering remote consultations and reducing the need for in-person visits. Meanwhile, in the field of ecotourism, destinations like national parks and protected areas provide a unique opportunity for travelers to explore natural ecosystems while contributing to conservation efforts.*"

contains terms from different domains, such as agriculture, healthcare, and ecotourism. KnowledgeTB successfully recognized and processed these domain-specific terms within the same text, demonstrating its ability to extract and enrich terminology from multiple domains simultaneously. Specifically, the algorithm detected terms and enhanced them from Wikidata from agriculture (e.g., *sustainable agriculture* (Q2751054), *carbon footprint* (Q310667), *crop rotation* (Q191258)), healthcare (e.g., *telemedicine* (Q46994)), and tourism (e.g., *ecotourism* (Q187449), *national parks* (Q28381982), *protected areas* (Q473972)). However, it was not able to link the term *national parks* because a Wikidata entry doesn't exist and missed recognizing general terms such as *patient care*, *patient*, *consultations*, *natural ecosystems*, and *ecosystem*. This highlights the need for further refinement in the system's sensitivity to accurately identify and link relevant terms.

The evaluation of KnowledgeTB's performance reveals a strong precision of 100%, indicating that all identified terms were relevant and correctly categorized into their respective domains. This highlights the system's ability to avoid false positives, which is crucial for maintaining the accuracy of domain-specific terminology. However, the recall stands at 50%, reflecting that only half of the expected terms were successfully identified. This indicates that while the system excels at correctly extracting terms, it occasionally misses relevant terminology, suggesting room for improvement in recall-oriented aspects. The balanced F1-score of 66.7% emphasizes the need to enhance recall while maintaining high precision. Additionally, the system achieved perfect domain accuracy of 100%, correctly assigning identified terms to their appropriate domains, demonstrating robustness in multi-domain scenarios. The enrichment success rate of 83.3% shows that the majority of identified terms were successfully linked to external knowledge sources, like Wikidata, though one term failed to link, pointing to potential gaps in linkage algorithms. Furthermore, the cross-domain accuracy of 50% indicates that while KnowledgeTB effectively processes multi-domain text, it still struggles with identifying and linking terms across diverse subject areas. Addressing the gaps in recall and improving cross-domain detection will further enhance the overall effectiveness and adaptability of the system.

Evaluation has been conducted on a limited dataset due to resource constraints. However, preliminary results indicate strong precision, and future work will extend testing to larger, diverse corpora to validate performance across domains.

## 5. Conclusion

KnowledgeTB shows strong potential as an innovative tool for terminology extraction and enrichment, offering a hybrid approach that integrates the power of LLMs and KGs. Preliminary results indicate that the system performs well in terms of extracting accurate terminology, linking terms to comprehensive metadata, and ensuring semantic enrichment through the use of Wikidata. Furthermore, the tool's multilingual capabilities enhance its global applicability, especially in domains like environmental science, where cross-border collaboration is essential.

While the system's testing has not been exhaustive, the early results demonstrate its ability to scale efficiently with large datasets, process domain-specific jargon, and provide enriched, contextually aware terminological resources. The combination of precise term extraction, semantic linking, and multilingual support positions KnowledgeTB as a valuable resource for building domain-specific glossaries, supporting semantic search, and enabling better knowledge representation.

Moving forward, further evaluation and refinement will focus on enhancing the tool's performance, particularly in specialized and highly technical domains. Additionally, feedback from users across a wider range of industries will help improve the system's usability and practicality. With continued development, KnowledgeTB has the potential to become a transformative tool for terminology management, facilitating better knowledge sharing, global collaboration, and decision-making in complex, knowledge-intensive fields.

While KnowledgeTB demonstrates strong precision, its reliance on Wikidata limits enrichment coverage for niche domains. Moreover, LLM outputs can vary depending on prompt phrasing. Addressing these challenges will require fine-tuning and broader knowledge graph integration in future work.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling checks. The authors have subsequently reviewed and edited the content and take full responsibility for the publication's final version.

## References

[1] R. Temmerman. *Towards New Ways of Terminology Description: The Sociocognitive Approach.* John Benjamins Publishing, 2000.

[2] K. Janowicz, P. Hitzler, W. Li, et al. KnowWhereGraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence. *AI Magazine*, vol. 43, pp. 30–39, 2022.

[3] OpenAI, J. Achiam, S. Adler, et al. GPT-4 Technical Report. *arXiv preprint* arXiv:2303.08774, 2024.

[4] H. Touvron, T. Lavril, G. Izacard, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint* arXiv:2302.13971, 2023.

[5] J. Giguere. Leveraging Large Language Models to Extract Terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pp. 57–60, Varna, Bulgaria, INCOMA Ltd., 2023.

[6] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint* arXiv:2403.08319, 2024.

[7] J. J. Norheim, E. Rebentisch, D. Xiao, L. Draeger, A. Kerbrat, and O. L. de Weck. Challenges in Applying Large Language Models to Requirements Engineering Tasks. Cambridge University Press, 2004.

[8] P. Cimiano, C. Chiarcos, J. P. McCrae, and J. Gracia. *Linguistic Linked Data: Representation, Generation, and Applications*. Springer, 2020.

[9] D. Vrandečić and M. Krötzsch. *Wikidata: A Free Collaborative Knowledgebase*. Communications of the ACM, vol. 57, no. 10, pp. 78–85, 2014.

[10] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne. Knowledge Graphs: Opportunities and Challenges. *arXiv preprint* arXiv:2303.13948, 2023.

[11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, vol. 3, article 160018, 2016.

[12] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[13] S. Kim, M. Sung, J. Lee, H. Lim, J. F. Gimenez Perez, "Efficient Terminology Integration for LLM-based Translation in Specialized Domains," 2023.

[14] S. Chataut, T. Do, B. D. S. Gurung, S. Aryal, A. Khanal, C. Lushbough, E. Gnimpieba, "Comparative Study of Domain Driven Terms Extraction Using Large Language Models," 2024.

[15] H. T. H. Tran, C.-E. González-Gallardo, J. Delaunay, A. Doucet, S. Pollak, "Is Prompting What Term Extraction Needs?". Text, Speech, and Dialogue: 27th International Conference, TSD 2024, Brno, Czech Republic, September 9–13, 2024, Proceedings, Part I. Pages 17 - 29. 2023.

[16] H. Babaei Giglou, J. D'Souza, S. Auer, "LLMs4OL: Large Language Models for Ontology Learning," 2023.

[17] J. P. McCrae, "Enriching a Terminology for Under-resourced Languages Using Knowledge Graphs". In Proceedings of eLex. 2021.

[18] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. *DBpedia: A Nucleus for a Web of Open Data*. The Semantic Web, vol. 4825, pp. 722–735, 2007.

[19] K. Ramakrishnan, G. Ramesh, and K. Sekar, "Trie: An Alternative Data Structure for Data Mining Algorithms," *Mathematical and Computer Modelling* , vol. 38(7-9):739-751, 2003.

[20] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, "The Falcon Series of Open Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2311.16867

[21] OpenAI, "GPT-3: Language Models are Few-Shot Learners," OpenAI, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.